Kernel Methods

- Generalization -

Hachem Kadri
hachem.kadri@lis-lab.fr

LIS, CNRS, Aix-Marseille Université, France

based on L. Ralaivola’s course on statistical learning theory

Outline

Focus: generalization bounds

Targetted result

$\mathcal{H}$ a family of function. $\forall \delta \in (0,1],$, with probability at least $1-\delta$ over the random draw of $S = \{(X_i, Y_i)\}_{i=1}^n$ the following holds

$$\forall h \in \mathcal{H}, \qquad \mathbb{E}_{XY} \ell(h, X, Y) \leq \frac{1}{n} \sum_{i=1}^{n} \ell(h, X_i, Y_i) + \varepsilon\left(\frac{1}{\delta}, \frac{1}{n}, \dots\right).$$

For binary classification we may want something like: with prob. $1-\delta$

$$\forall h \in \mathcal{H}, \mathbb{P}_{XY}(h(X) \neq Y) \leq \hat{R}(h, S) + \varepsilon\left(\frac{1}{\delta}, \frac{1}{n}, \dots\right).$$

Remark (On ε)

- decreases when n increases and when δ increases
- usually contains something related to the *capacity* of $\mathcal{H}$

Focus: generalization bounds

Targetted result

$\mathcal{H}$ a family of function. $\forall \delta \in (0,1],$, with probability at least $1-\delta$ over the random draw of $S = \{(X_i, Y_i)\}_{i=1}^n$ the following holds

$$\forall h \in \mathcal{H}, \qquad \mathbb{E}_{XY}\ell(h, X, Y) \leq \frac{1}{n}\sum_{i=1}^{n} \ell(h, X_i, Y_i) + \varepsilon\left(\frac{1}{\delta}, \frac{1}{n}, \ldots\right).$$

For binary classification we may want something like: with prob. $1-\delta$

$$\forall h \in \mathcal{H}, \mathbb{P}_{XY}(h(X) \neq Y) \leq \hat{R}(h, S) + \varepsilon\left(\frac{1}{\delta}, \frac{1}{n}, \ldots\right).$$

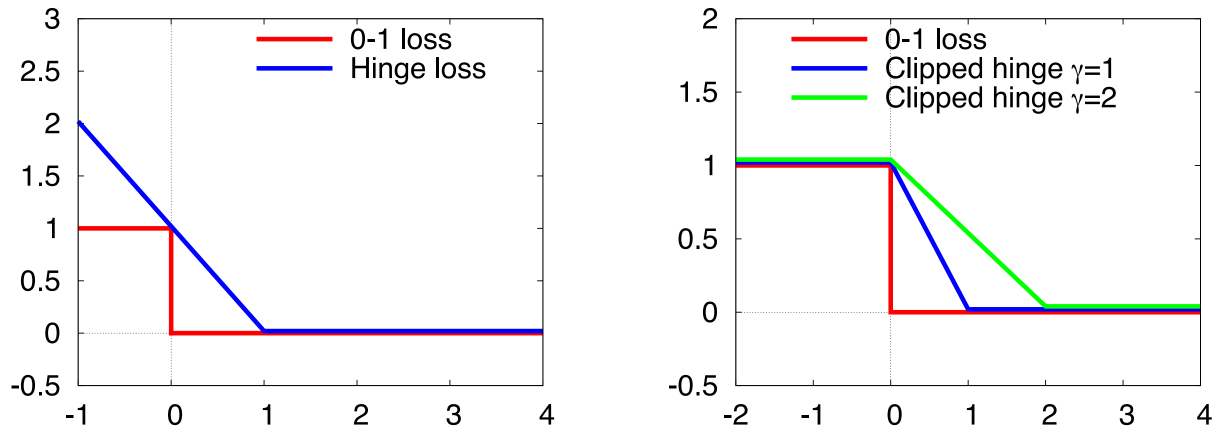
Remark

Many ways to get generalization bounds

- VC dimension-based arguments [Vapnik, 1998]
- PAC-Bayesian theory [McAllester, 1999]
- Algorithmic stability theory [Bousquet and Elisseeff, 2002]
- Rademacher-complexity based arguments (our focus) [Bartlett and Mendelson, 2002]
- ...

Generalization bounds for binary kernel classifiers (e.g. SVMs)

0-1 loss and upper bounds

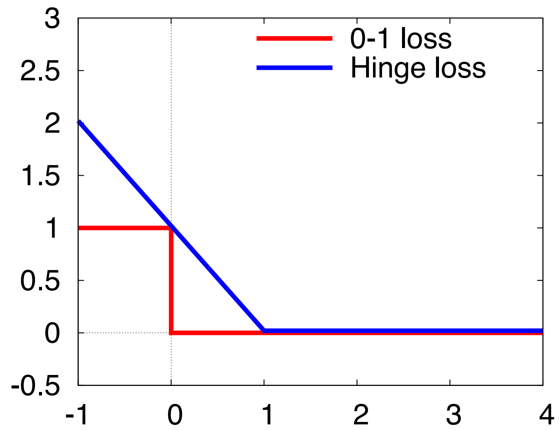


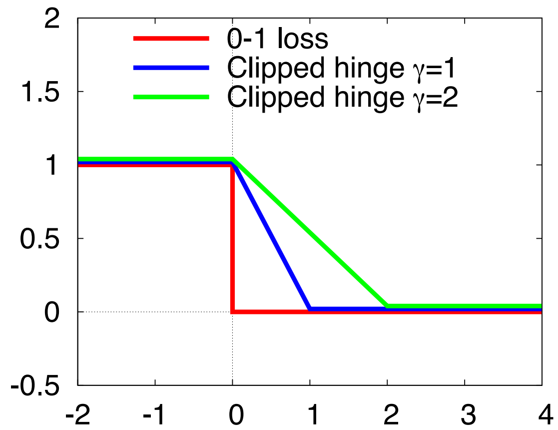
Clipped loss ℓ^c_γ

$$\ell^c_\gamma(\theta) = \begin{cases} 1 & \text{if } \theta \leq 0 \\ 1 - \theta/\gamma & \text{if } 0 < \theta \leq \gamma \\ 0 & \text{otherwise.} \end{cases}$$

Generalization bounds for binary kernel classifiers (e.g. SVMs)

0-1 loss and upper bounds





Remark (On the clipped hinge loss)

- Upper bound on the 0-1 binary loss: $\forall \theta,\ 0 \leq \ell_{0-1}(\theta) \leq \ell_{\gamma}^{c}(\theta)$
- The empirical clipped $\hat{R}_{\ell_{\gamma}^{c}}(h, S)$ risk is $\hat{R}_{\ell_{\gamma}^{c}}(h, S) = \dfrac{1}{n} \sum_{i=1}^{n} \ell_{\gamma}^{c}(Y_i h(X_i))$
- It is $1/\gamma$-Lipschitz:

$$\forall \theta, \theta',\ |\ell_{\gamma}^{c}(\theta) - \ell_{\gamma}^{c}(\theta')\| \leq \frac{1}{\gamma} |\theta - \theta'|$$

Generalization bounds for binary kernel classifiers (e.g. SVMs)

Assumptions

- k, a bounded Mercer kernel: $\sup k(X, X) \leq R^2$
- Bounded norm functions

$$\mathcal{H}_\Lambda = \left\{ \mathbf{x} \mapsto \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}),\ \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n,\ \|f\|^2 = \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \leq \Lambda^2 \right\}$$

Theorem (Rademacher generalization bound for kernel classifiers [Bartlett and Mendelson, 2002, Shawe-Taylor and Cristianini, 2004])

$\forall \delta \in [0, 1)$, *with probability at least* $1 - \delta$, $\forall h \in \mathcal{H}_\Lambda$,

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \underbrace{\frac{1}{n} \sum_{i=1}^n \ell_\gamma^c(Y_i h(X_i))}_{\hat{R}_{\ell_\gamma^c}(h,S)} + \frac{c_1 \Lambda}{\gamma n} \sqrt{\sum_{i=1}^n k(X_i, X_i)} + c_2 \sqrt{\frac{\ln 4/\delta}{2n}}$$

where $c_1, c_2 > 0$.

Generalization bounds for binary kernel classifiers (e.g. SVMs)

Assumptions

- k, a bounded Mercer kernel: $\sup k(X, X) \leq R^2$
- Bounded norm functions

$$\mathcal{H}_\Lambda = \left\{ \mathbf{x} \mapsto \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}),\ \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X}^n,\ \|f\|^2 = \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \leq \Lambda^2 \right\}$$

Remark

- *Data-dependent* generalization bound
- Since $\sqrt{\sum_{i=1}^n k(X_i, X_i)} \leq \Lambda\sqrt{n}$, at least a $O(1/\sqrt{n})$ decreasing rate

Diving into the proof (cooooooool !!)

Theorem ([McDiarmid, 1989])

Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $\mathcal{X}$. Assume that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_i' \in \mathcal{X}} \left| f(\mathbf{x}_1, \ldots, \mathbf{x}_n) - f(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_i', \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n) \right| \leq c_i$$

for every $1 \leq i \leq n$. Then, for every $t > 0$,

$$P\left\{|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)| \geq t\right\} \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Remark

A generalization of Chernoff-Hoeffding bounds. Indeed, if $f(\mathbf{x}_1, \ldots \mathbf{x}_n) = \frac{1}{n}\sum_{i=1}^n x_i$ and $X_i \in [a, b]$ and IID, then

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_n,\mathbf{x}_i' \in \mathcal{X}} \left| f(\mathbf{x}_1, \ldots, \mathbf{x}_n) - f(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_i', \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n) \right| \leq \frac{|b-a|}{n}$$

and

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X\right| \geq t\right\} \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Diving into the proof (cooooooool !!)

$$\mathbb{P}_{XY}(Yh(X) \leq 0) = \mathbb{E}_{XY}\ell_{0-1}(Yh(X))$$

Diving into the proof (cooooooool !!)

$$\begin{aligned}\mathbb{P}_{XY}(Yh(X) \leq 0) &= \mathbb{E}_{XY}\ell_{0-1}(Yh(X)) \\ &\leq \mathbb{E}_{XY}\ell_{\gamma}^{c}(Yh(X))\end{aligned}$$

Diving into the proof (coooooool !!)

$$
\begin{aligned}
\mathbb{P}_{XY}(Yh(X) \le 0) &= \mathbb{E}_{XY}\ell_{0-1}(Yh(X)) \\
&\le \mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) \\
&= \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i)) + \mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i))
\end{aligned}
$$

Diving into the proof (cooooooool !!)

$$
\begin{aligned}
\mathbb{P}_{XY}(Yh(X) \leq 0) &= \mathbb{E}_{XY}\ell_{0-1}(Yh(X)) \\
&\leq \mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) \\
&= \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i)) + \mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i)) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i)) + \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i))\right]
\end{aligned}
$$

Diving into the proof (cooooooool !!)

$$
\begin{aligned}
\mathbb{P}_{XY}(Yh(X) \leq 0) &= \mathbb{E}_{XY}\ell_{0-1}(Yh(X)) \\
&\leq \mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) \\
&= \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i)) + \mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i)) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i)) + \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i))\right] \\
&= \hat{R}_{\ell_\gamma^c}(h, S) + \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \frac{1}{n}\sum_{i=1}^{n}\ell_\gamma^c(Y_i h(X_i))\right]
\end{aligned}
$$

Diving into the proof (cooooooooool !!!)

Let us take care of $\sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell^c_\gamma(Yh(X)) - \frac{1}{n}\sum_{i=1}^n \ell^c_\gamma(Y_i h(X_i))\right]$:
The function

$$H((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)) := \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell^c_\gamma(Yh(X)) - \frac{1}{n}\sum_{i=1}^n \ell^c_\gamma(y_i h(\mathbf{x}_i))\right]$$

is such that

$$\sup_{(\mathbf{x}_i, y_i), (\mathbf{x}'_i, y'_i)} \left|H((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)) - H((\mathbf{x}_1, y_1), \dots, (\mathbf{x}'_i, y'_i), \dots, (\mathbf{x}_n, y_n))\right| \leq \frac{1}{n}$$

Hence, using McDiarmid's inequality

$$\mathbb{P}_S\left(\left|\mathbb{E}_{S'}H(S') - H(S)\right| \geq t\right) \leq 2\exp\left(-2nt^2\right)$$

Or, solving for $2\exp\left(-2nt^2\right) = \delta$, with prob. at least $1-\delta$

$$H(S) \leq \mathbb{E}_{S'}H(S') + \sqrt{\frac{\log 2/\delta}{2n}}$$

Diving into the proof (coooooooool !!!)

Let us take care of $\sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \frac{1}{n}\sum_{i=1}^n \ell_\gamma^c(Y_i h(X_i))\right]$:
The function

$$H((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)) := \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \frac{1}{n}\sum_{i=1}^n \ell_\gamma^c(y_i h(\mathbf{x}_i))\right]$$

is such that

$$\sup_{(\mathbf{x}_i,y_i),(\mathbf{x}_i',y_i')}\left|H((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)) - H((\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_i',y_i'),\ldots,(\mathbf{x}_n,y_n))\right| \leq \frac{1}{n}$$

Hence, using McDiarmid's inequality

$$\mathbb{P}_S\left(\left|\mathbb{E}_{S'}H(S') - H(S)\right| \geq t\right) \leq 2\exp\left(-2nt^2\right)$$

In other words, with probability at least $1-\delta$

$$\sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \frac{1}{n}\sum_{i=1}^n \ell_\gamma^c(Y_i h(X_i))\right]$$
$$\leq \mathbb{E}_S \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \hat{R}_{\ell_\gamma^c}(h,S)\right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Diving into the proof (coooooooooool !!!)

Going back where we left off, with probability at least $1 - \delta$

$$\mathbb{P}_{XY}(Yh(X) \le 0) \le \hat{R}_{\ell^c_\gamma}(h, S) + \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h, S) \right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Diving into the proof (cooooooooool !!!)

Going back where we left off, with probability at least $1-\delta$

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell^c_\gamma}(h,S) + \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h,S) \right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Now, let us deal with the second term of the right-hand side:

$$\begin{aligned}
&\mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h,S) \right] \\
&\quad = \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{S'} \hat{R}_{\ell^c_\gamma}(h,S') - \hat{R}_{\ell^c_\gamma}(h,S) \right]
\end{aligned}$$

Diving into the proof (cooooooooool !!!)

Going back where we left off, with probability at least $1-\delta$

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell_\gamma^c}(h,S) + \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell_\gamma^c(Yh(X)) - \hat{R}_{\ell_\gamma^c}(h,S) \right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Now, let us deal with the second term of the right-hand side:

$$\begin{aligned}
&\mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell_\gamma^c(Yh(X)) - \hat{R}_{\ell_\gamma^c}(h,S) \right] \\
&\quad = \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{S'} \hat{R}_{\ell_\gamma^c}(h,S') - \hat{R}_{\ell_\gamma^c}(h,S) \right] \\
&\quad = \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{S'} \left(\hat{R}_{\ell_\gamma^c}(h,S') - \hat{R}_{\ell_\gamma^c}(h,S) \right) \right]
\end{aligned}$$

Diving into the proof (coooooooool !!!)

Going back where we left off, with probability at least $1-\delta$

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell_\gamma^c}(h,S) + \mathbb{E}_S \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \hat{R}_{\ell_\gamma^c}(h,S)\right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Now, let us deal with the second term of the right-hand side:

$$\begin{aligned}
&\mathbb{E}_S \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell_\gamma^c(Yh(X)) - \hat{R}_{\ell_\gamma^c}(h,S)\right] \\
&= \mathbb{E}_S \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{S'}\hat{R}_{\ell_\gamma^c}(h,S') - \hat{R}_{\ell_\gamma^c}(h,S)\right] \\
&= \mathbb{E}_S \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{S'}\left(\hat{R}_{\ell_\gamma^c}(h,S') - \hat{R}_{\ell_\gamma^c}(h,S)\right)\right] \\
&\leq \mathbb{E}_{SS'} \sup_{h\in\mathcal{H}_\Lambda}\left[\frac{1}{n}\sum_{i=1}^n \ell_\gamma^c(Y_i'h(X_i')) - \frac{1}{n}\sum_{i=1}^n \ell_\gamma^c(Y_ih(X_i))\right]
\end{aligned}$$

because sup is convex (this is *Jensen's inequality* applied to this function)

Diving into the proof (cooooooooool !!!)

Going back where we left off, with probability at least $1-\delta$

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell_\gamma^c}(h,S) + \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell_\gamma^c(Yh(X)) - \hat{R}_{\ell_\gamma^c}(h,S)\right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Now, let us deal with the second term of the right-hand side:

$$\begin{aligned}
&\mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell_\gamma^c(Yh(X)) - \hat{R}_{\ell_\gamma^c}(h,S)\right] \\
&\qquad = \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{S'} \hat{R}_{\ell_\gamma^c}(h,S') - \hat{R}_{\ell_\gamma^c}(h,S)\right] \\
&\qquad = \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{S'} \left(\hat{R}_{\ell_\gamma^c}(h,S') - \hat{R}_{\ell_\gamma^c}(h,S)\right)\right] \\
&\qquad \leq \mathbb{E}_{SS'} \sup_{h \in \mathcal{H}_\Lambda} \left[\frac{1}{n}\sum_{i=1}^n \ell_\gamma^c(Y_i' h(X_i')) - \frac{1}{n}\sum_{i=1}^n \ell_\gamma^c(Y_i h(X_i))\right]
\end{aligned}$$

because sup is convex (this is *Jensen's inequality* applied to this function)

$$= \mathbb{E}_{SS'} \sup_{h \in \mathcal{H}_\Lambda} \left[\frac{1}{n}\sum_{i=1}^n \left(\ell_\gamma^c(Y_i' h(X_i')) - \ell_\gamma^c(Y_i h(X_i))\right)\right]$$

Diving into the proof (coooooooool !!!)

Going back where we left off, with probability at least $1 - \delta$

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell^c_\gamma}(h, S) + \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h, S) \right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Take a deep breath, and let us keep dealing with this second term

$$\begin{aligned}
&\mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h, S) \right] \\
&\qquad = \mathbb{E}_{SS'} \sup_{h \in \mathcal{H}_\Lambda} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\ell^c_\gamma(Y_i' h(X_i')) - \ell^c_\gamma(Y_i h(X_i)) \right) \right]
\end{aligned}$$

Diving into the proof (cooooooooool !!!)

Going back where we left off, with probability at least $1-\delta$

$$\mathbb{P}_{XY}(Yh(X)\leq 0) \leq \hat{R}_{\ell^c_\gamma}(h,S) + \mathbb{E}_S \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h,S)\right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Take a deep breath, and let us keep dealing with this second term

$$\begin{aligned}
&\mathbb{E}_S \sup_{h\in\mathcal{H}_\Lambda}\left[\mathbb{E}_{XY}\ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h,S)\right] \\
&\qquad = \mathbb{E}_{SS'} \sup_{h\in\mathcal{H}_\Lambda}\left[\frac{1}{n}\sum_{i=1}^n \left(\ell^c_\gamma(Y_i'h(X_i')) - \ell^c_\gamma(Y_ih(X_i))\right)\right] \\
&\qquad = \mathbb{E}_{SS'\underline{\sigma}} \sup_{h\in\mathcal{H}_\Lambda}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i \left(\ell^c_\gamma(Y_i'h(X_i')) - \ell^c_\gamma(Y_ih(X_i))\right)\right]
\end{aligned}$$

where $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$ and we used that S and S' are IID

Diving into the proof (cooooooooool !!!)

Going back where we left off, with probability at least $1 - \delta$

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell^c_\gamma}(h, S) + \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h, S) \right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Take a deep breath, and let us keep dealing with this second term

$$\begin{aligned}
\mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} & \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h, S) \right] \\
&= \mathbb{E}_{SS'} \sup_{h \in \mathcal{H}_\Lambda} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\ell^c_\gamma(Y'_i h(X'_i)) - \ell^c_\gamma(Y_i h(X_i)) \right) \right] \\
&= \mathbb{E}_{SS'\underline{\sigma}} \sup_{h \in \mathcal{H}_\Lambda} \left[\frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\ell^c_\gamma(Y'_i h(X'_i)) - \ell^c_\gamma(Y_i h(X_i)) \right) \right]
\end{aligned}$$

where $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$ and we used that S and S' are IID

$$\leq \mathbb{E}_{SS'\underline{\sigma\sigma'}} \sup_{h \in \mathcal{H}_\Lambda} \left[\frac{1}{n} \sum_{i=1}^{n} \sigma'_i \ell^c_\gamma(Y'_i h(X'_i)) - \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell^c_\gamma(Y_i h(X_i)) \right]$$

Diving into the proof (cooooooooool !!!)

Going back where we left off, with probability at least $1-\delta$

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell^c_\gamma}(h,S) + \mathbb{E}_S \sup_{h\in\mathcal{H}_\Lambda} \left[\mathbb{E}_{XY}\ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h,S)\right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

Take a deep breath, and let us keep dealing with this second term

$$\begin{aligned}
&\mathbb{E}_S \sup_{h\in\mathcal{H}_\Lambda} \left[\mathbb{E}_{XY}\ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h,S)\right] \\
&\qquad = \mathbb{E}_{SS'} \sup_{h\in\mathcal{H}_\Lambda} \left[\frac{1}{n}\sum_{i=1}^{n}\left(\ell^c_\gamma(Y_i'h(X_i')) - \ell^c_\gamma(Y_ih(X_i))\right)\right] \\
&\qquad = \mathbb{E}_{SS'\underline{\sigma}} \sup_{h\in\mathcal{H}_\Lambda} \left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\ell^c_\gamma(Y_i'h(X_i')) - \ell^c_\gamma(Y_ih(X_i))\right)\right]
\end{aligned}$$

where $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$ and we used that S and S' are IID

$$\begin{aligned}
&\leq \mathbb{E}_{SS'\underline{\sigma\sigma'}} \sup_{h\in\mathcal{H}_\Lambda} \left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i'\ell^c_\gamma(Y_i'h(X_i')) - \frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell^c_\gamma(Y_ih(X_i))\right] \\
&\leq \mathbb{E}_{S\underline{\sigma}}\frac{2}{n} \sup_{h\in\mathcal{H}_\Lambda} \left|\sum_{i=1}^{n}\sigma_i\ell^c_\gamma(Y_ih(X_i))\right|
\end{aligned}$$

Diving into the proof (cooooooooool !!!)

Going back where we left off, with probability at least $1 - \delta$

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell^c_\gamma}(h,S) + \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h,S) \right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

This second term is resilient, but we are almost there

$$\mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h,S) \right] \leq \mathbb{E}_{S\underline{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}_\Lambda} \left| \sum_{i=1}^n \sigma_i \ell^c_\gamma(Y_i h(X_i)) \right|$$

Diving into the proof (cooooooooool !!!)

Going back where we left off, with probability at least $1 - \delta$

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell^c_\gamma}(h, S) + \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h, S) \right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

This second term is resilient, but we are almost there

$$\mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell^c_\gamma(Yh(X)) - \hat{R}_{\ell^c_\gamma}(h, S) \right] \leq \mathbb{E}_{S\underline{\sigma}} \frac{2}{n} \sup_{h \in \mathcal{H}_\Lambda} \left| \sum_{i=1}^{n} \sigma_i \ell^c_\gamma(Y_i h(X_i)) \right|$$

$$\leq \underbrace{\frac{2}{\gamma} \mathbb{E}_{S\underline{\sigma}} \sup_{h \in \mathcal{H}_\Lambda} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i h(X_i) \right|}_{\text{see [Bartlett and Mendelson, 2002]}}$$

Rademacher complexity of $\mathcal{H}_\Lambda$

Definition

$$\mathcal{R}(\mathcal{H}_\Lambda) = \mathbb{E}_{S\underline{\sigma}} \sup_{h \in \mathcal{H}_\Lambda} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i h(X_i) \right|$$

is the *Rademacher complexity* of $\mathcal{H}_\Lambda$

Remark

- It measures the richness of the class $\mathcal{H}_\Lambda$
- Based on how well the class of functions is capable of correlating with randomly assigned labels
- The marginal distribution over $\mathcal{X}$ is directly taken into account

Definition (Empirical Rademacher complexity $\hat{\mathcal{R}}(\mathcal{H}_\Lambda, S)$)

The *empirical Rademacher Complexity* of $\mathcal{H}_\Lambda$ is defined as

$$\hat{\mathcal{R}}(\mathcal{H}_\Lambda, S) := \mathbb{E}_{\underline{\sigma}} \sup_{h \in \mathcal{H}_\Lambda} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i h(X_i) \right|$$

Rademacher complexity of $\mathcal{H}_\Lambda$

Concentration of $\hat{\mathcal{R}}(\mathcal{H}_\Lambda, S)$

$\hat{\mathcal{R}}(\mathcal{H}_\Lambda, S)$ is a concentrated variable. Using McDiarmid inequality again, we obtain that, with probability at least $1 - \delta$

$$\mathcal{R}(\mathcal{H}_\Lambda) \leq \hat{\mathcal{R}}(\mathcal{H}_\Lambda, S) + c\sqrt{\frac{\log 2/\delta}{2n}}$$

Bounding $\hat{\mathcal{R}}(\mathcal{H}_\Lambda, S)$

$$\begin{aligned}
\hat{\mathcal{R}}(\mathcal{H}_\Lambda, S) &= \mathbb{E}_{\underline{\sigma}} \sup_{h \in \mathcal{H}_\Lambda} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i h(X_i) \right| = \mathbb{E}_{\underline{\sigma}} \sup_{h \in \mathcal{H}_\Lambda} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i \langle h, k(X_i, \cdot) \rangle \right| \\
&= \mathbb{E}_{\underline{\sigma}} \sup_{h \in \mathcal{H}_\Lambda} \frac{2}{n} \left| \left\langle h, \sum_{i=1}^{n} \sigma_i k(X_i, \cdot) \right\rangle \right| \\
&= \mathbb{E}_{\underline{\sigma}} \frac{2}{n} \Lambda \left\| \sum_{i=1}^{n} \sigma_i k(X_i, \cdot) \right\| = \mathbb{E}_{\underline{\sigma}} \frac{2}{n} \Lambda \sqrt{\sum_{ij} \sigma_i \sigma_j k(X_i, X_j)} \\
&\leq \frac{2}{n} \Lambda \sqrt{\mathbb{E}_{\underline{\sigma}} \sum_{ij} \sigma_i \sigma_j k(X_i, X_j)} = \frac{2}{n} \Lambda \sqrt{\sum_{i} k(X_i, X_i)}
\end{aligned}$$

Closing the proof

We had

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell_\gamma^c}(h, S) + \mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell_\gamma^c(Yh(X)) - \hat{R}_{\ell_\gamma^c}(h, S) \right] + \sqrt{\frac{\log 2/\delta}{2n}}$$

and

$$\mathbb{E}_S \sup_{h \in \mathcal{H}_\Lambda} \left[\mathbb{E}_{XY} \ell_\gamma^c(Yh(X)) - \hat{R}_{\ell_\gamma^c}(h, S) \right] \leq \frac{2}{\gamma} \mathbb{E}_{S\underline{\sigma}} \sup_{h \in \mathcal{H}_\Lambda} \frac{2}{n} \left| \sum_{i=1}^n \sigma_i h(X_i) \right|$$

and we just proved

$$\mathbb{E}_{S\underline{\sigma}} \sup_{h \in \mathcal{H}_\Lambda} \frac{2}{n} \left| \sum_{i=1}^n \sigma_i h(X_i) \right| \leq \frac{2}{n} \Lambda \sqrt{\sum_i k(X_i, X_i)} + c \sqrt{\frac{\log 2/\delta}{2n}}$$

Combining everything and adjusting some constants give the desired result

$$\mathbb{P}_{XY}(Yh(X) \leq 0) \leq \hat{R}_{\ell_\gamma^c}(h, S) + \frac{c_1 \Lambda}{\gamma n} \sqrt{\sum_{i=1}^n k(X_i, X_i)} + c_2 \sqrt{\frac{\ln 4/\delta}{2n}}$$

Partial conclusion

Generalization bounds

- are easy to derive for kernel classifiers
- can still be improved
- need for practical bounds (such as leave-one-out bounds)
- . . .

References I

Bartlett, P. L. and Mendelson, S. (2002).
Rademacher and gaussian complexities: Risk bounds and structural results.
Journal of Machine Learning Research, 3:463–482.

Bousquet, O. and Elisseeff, A. (2002).
Stability and Generalization.
Journal of Machine Learning Research, 2:499–526.

McAllester, D. (1999).
Pac-bayesian model averaging.
In *Proc. of the 12th Annual Conf. on Comp. learning theory*, pages 164–170, New York, NY, USA.

McDiarmid, C. (1989).
On the method of bounded differences.
Survey in Combinatorics, pages 148–188.

Shawe-Taylor, J. and Cristianini, N. (2004).
Kernel Methods for Pattern Analysis.
Cambridge University Press.

References II

Vapnik, V. (1998).
Statistical Learning Theory.
John Wiley and Sons, inc.